OPEN ACCESS

Journal of Oncology Research and Therapy

GAVIN PUBLISHERS

Research Article

Karagiannis E, et al. J Oncol Res Ther 6: 10113.

DOI: 10.29011/2574-710X.100113

Evaluation of an Atlas-Based Auto-Segmentation Tool of Target Volumes and Organs at Risk in Head and Neck Radiation Therapy

Efstratios Karagiannis[1,2*], Panagiotis Koreas[3], Iosif Strouthos[1,2], Agnes Leczynski[1], Marcus Grimm[4], Nikolaos Zamboglou[1,2], Konstantinos Ferentinos[1,2]

[1]Department of Radiation Oncology, German Oncology Center, Limassol, Cyprus

[2]Department of Medicine, School of Medicine, European University Cyprus, Nicosia, Cyprus

[3]Faculty of Medicine, "Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, Cluj-Napoca, Romania

[*]**Corresponding author:** Efstratios Karagiannis, Department of Radiation Oncology, German Oncology Centre 1 Nikis Avenue, Limassol, Cyprus.

Citation: Karagiannis E, Koreas P, Strouthos I, Leczynski A, Grimm M, et al. (2021) Evaluation of an Atlas-Based Auto-Segmentation Tool of Target Volumes and Organs at Risk in Head and Neck Radiation Therapy. J Oncol Res Ther 6: 10113. DOI: 10.29011/2574-710X.100113

Received Date: 01 September, 2021; **Accepted Date:** 13 September, 2021; **Published Date**: 17 September, 2021

Abstract

Introduction: Radiation therapy plays a crucial role in the required multidisciplinary management of Head And Neck Cancer (HNC). The use of highly sophisticated radiation therapy techniques for HNC mandates highly precise target volumes and Organs at Risk (OARs) delineation, which can be challenging. Computerised contouring algorithms have shown potential by improving delineation precision, inter-observer consistency and reducing workload. This study aims to evaluate a commercially available Atlas-Based-Auto-Segmentation (ABAS) module, assessing its accuracy, clinical applicability, as well as to explore the potential role of ABAS in the evolving era of artificial intelligence and deep learning contouring (DLC).

Methods: The ABAS model was created using 100 HNC patients' imaging data. A second cohort of 20 patients imaging data was used to evaluate the ABAS delineation results compared to expert manual contours. For each of the 39 regions of interest (ROIs) for every patient, commonly used quantitative metrics, the Dice Similarity Coefficient Index (DICE), the Hausdorff distance 95th-percentile (HD) and the Volume Ratio (VR) were obtained and compared. A subjective evaluation (Turing test) and a time evaluation were also performed.

Results: The performance of the ABAS model tested, regarding the quantitative metrics, was similar to other ABAS solutions, and mostly, slightly inferior to presented DLC solutions. The subjective evaluation showed that although not all ABAS contours are precise, it is not always straightforward to identify contours as being human- or computer-created. The time evaluation suggested that ABAS, followed by manual editing, can significantly reduce the time and effort needed for the segmentation of the ROIs of the head and neck regions.

Conclusion: Auto contouring will play an important role in the future of radiotherapy planning. Using ABAS or DLC modules and fine-tuning the results is efficient for clinical utility and can considerably save time for clinicians in delineating ROIs for the head and neck region.

Keywords: Auto-Contouring; Auto-Segmentation; Head and Neck Cancer; Radiation Therapy

Introduction

Head and Neck Cancer (HNC) represents a global public health issue with a significant socioeconomic impact, as it accounts for more than 650,000 new cases and 300,000 deaths annually, not including thyroid cancer [1]. It is the sixth most common cancer in the world and affects significantly more males than females, with a ratio ranging from 2:1 to 4:1 [1-3]. Tobacco use (both smoked and smokeless), areca nut, betel quid, alcohol, and human papillomavirus (HPV) infection are considered to be the most well-studied risk factors for all types of HNC, but more specifically for oral and oropharyngeal cancers [4-6].

The optimal management for patients with HNC requires a multidisciplinary approach, as precision patient-specific medicine needs to be incorporated. The multidisciplinary team includes surgeons (oral and maxillofacial, otolaryngology, plastic and reconstructive), radiation oncologists, medical oncologists, radiologists, pathologists and nuclear medicine consultants, among others. Surgery and/or radiotherapy, with or without neo-adjuvant, simultaneous or adjuvant systemic therapy, are the key treatment modalities. The goal of radiotherapy is to eradicate cancer cells, mostly by damaging their DNA, while minimizing radiation damage to critical healthy tissue. Radiotherapy, utilising photons or protons, for patients with HNC is complex and has greatly evolved in the past decades, owing to the advent of intensity-modulated radiotherapy techniques and adaptive treatment planning.

Intensity-Modulated Radiation Therapy (IMRT) is an advanced form of high-precision, three-dimensional conformal radiotherapy, using computer-optimized inverse treatment planning and a computer-controlled multileaf collimator. With the use of static or rotational IMRT, the intensity of radiation beams can be modulated, so that a higher radiation dose can be delivered to the target volumes, with a sharply conformal coverage, while at the same time the dose to surrounding normal tissues is significantly reduced [7]. As IMRT is a highly precise technique, it requires apart from reproducible patient setup/immobilisation and optimal imaging modalities, highly precise target volumes and organs at risk (OARs) delineation.

Although delineation guidelines that contribute to the reduction of contouring variations and the simplification of multi-institutional clinical trials conduction, as well as to improved care of HNC patients, do exist [8], considerable variation in practice policy is noticed among clinicians and institutions [9]. Additionally, the average time required for adequate manual delineation of target volumes and OARs for HNC patients is approximately 120 minutes. Moreover, the already challenging task of manual contouring is getting more complex and time-demanding, as a consequence of the increasing number of OARs found to be associated with radiation-induced side effects. Against this background, auto-contouring of target volumes and OARs has the potential to reduce time and effort, and to improve delineation precision and inter-observer consistency [10-12].

Atlas-Based-Auto-Segmentation (ABAS) is a widely used method in which a set of a specific number of representative patients with carefully delineated OARs serves as a reference set for contouring new patients [13]. Contours of the reference patients are registered to new patients, in order to be transformed and implemented in the new scan. Although ABAS has already significantly reduced workload and improved consistency in clinical practice, there are a number of limitations, such as suboptimal performance for small target volumes and OARs or for patients with differing anatomies from those used as reference, as well as errors due to deformation inaccuracies [14-16].

Segmentation methods, using deep learning techniques have emerged as a promising method of coping with these challenges. Deep Learning Contouring (DLC) typically trains a Convolutional Neural Network (CNN) model directly from high quality data [17]. Improved computing power and adequate training of neural networks have made deep learning methods available for autosegmenation purposes. Several studies have already shown the potential of CNNs for HNC contouring and for other sites [18-22]. The key feature of the CNNs is their ability to recognise patterns, group and use them to predict results, without the need of human interventions, but more importantly, to be able to evolve on the basis of experience. It is self-explanatory that the CNNs performance is based on the quality of data used to train them. As noted, high quality data (manually produced) are challenging and time-demanding. ABAS-based, manually fine-tuned data might contribute to this task [17].

This study aims to evaluate a commercial ABAS module included into a dedicated radiotherapy contouring software (ProSoma, Medcom GmbH, Darmstadt, Germany) for HNC patients. The primary goal is to assess the accuracy and clinical applicability of the generated target and OARs contours, using quantitative geometric metrics and subjective evaluations, and to compare the time needed between manual and ABAS segmentation. The secondary goal of the study is to calculate the time for required manual corrections for the contours to be clinically acceptable, so that the software could be a useful tool in creating quality data for CNNs.

Materials and Methods

ProSoma atlas data set

A total of 100 patients' DICOM files were used to build the atlas data set. Eligible patients for this data set were patients with non-specific primary tumour or treatment site, that did not undergo a head-neck surgery for any reason, positioned and simulated in supine position, in order to be treated with curative or palliative radiation therapy. All the patients were immobilised with a custom five-point commercial thermoplastic mask (IMRT-Mask-Precut-MR-09^{Ò}, Unger Medizintechnik GmbH, Mülheim-Kärlich, Germany), using standard head-neck cushions (4-piece-set-of-Head-Support-CushionsÒ, Unger Medizintechnik GmbH, Mülheim-Kärlich, Germany).

For each patient a planning-CT scan (3 mm), using a multi-slice CT Siemens SOMATOM Sensation open (Siemens AG, Munich, Germany), was performed, followed by the manual segmentation of the target volumes and OARs. Target volumes included the bilateral lymphatic levels I-V, while the OARs included the bilateral lacrimal glands, eyes, lenses optic nerves, inner ears, parotids, submandibular glands, mandible (temporomandibular) joints, sternocleidomastoids, as well as the brain, brainstem, optic chiasma, hypophysis, spinal cord, larynx, hyoid, mandible, trachea and body surface. All the manual segmentations were delineated by a dedicated team of radiation oncologists with expertise in HNC, according to previously published international consensus delineation guidelines [23,24].

Reference and test data set

A total of 20 patients' DICOM files were used to build the reference and test data set. Eligible patients for this data set were HNC patients that did not undergo a head-neck surgery for any reason, positioned and simulated in supine position, in order to be treated with definitive radio chemotherapy. All the patients, similarly to the reference data set, were immobilised with a custom five-point commercial thermoplastic mask (IMRT-Mask-Precut-MR-09$^{\text{Ò}}$, Unger Medizintechnik GmbH, Mülheim-Kärlich, Germany), using standard head-neck cushions (4-piece-set-of-Head-Support-Cushions$^{\text{Ò}}$, Unger Medizintechnik GmbH, Mülheim-Kärlich, Germany).

For each patient, a planning-CT scan (2.5 mm), using a multi-slice CT GE Discovery CT590 RT (GE Healthcare, Chicago, USA), was performed, followed by the manual segmentation of the same target volumes and OARs as for the ProSoma atlas set. All the manual segmentations were delineated by a dedicated team of radiation oncologists with expertise in HNC, according to previously published international consensus delineation guidelines [23,24].

Automatic segmentation was subsequently performed for all patients using the auto contouring module of the software. For the automatic segmentation, the same target volumes and OARs as for the ProSoma atlas and the ABAS reference data sets were delineated.

Quantitative evaluation

The autocontouring module's performance was evaluated by comparing the differences between the automatically generated (ABAS test data set) and manual contours (ABAS reference data set) using the following metrics:

- the Dice similarity coefficient (DICE), which quantifies the overlap between contours A and B: $\mathrm{DICE} = \frac{2(A \cap B)}{A + B}$ [25]
- The Hausdorff distance 95th-percentile (HD), i.e. the 95th percentile of the pairwise 3D point distances between two structures' contour in mm [26].
- The volume ratio (VR) between the contours of the two groups

Subjective evaluation

A subjective evaluation of the contouring methods was carried out with a Turing test, which assumes clinical usability of auto-generated contours, if they are difficult to distinguish from manual contours [27]. A HNC radiation oncology expert (observer), who was not involved into the project, was invited to take the Turing test to prevent potential bias. Following the approach described by Gooding, et al. [27], the observer was blindly presented with random slices that had an equal probability of featuring manual or auto-generated contours for all target volumes and OARs. The observer assessed the following questions for 100 scenarios:

- A single contour: "How was this contour drawn?"

Answer options: "By a human" or "By a computer".

- Two contours: "Which contour do you prefer?"

The preferred contour is selected by the observer.

- A single contour: "You have been asked to evaluate this contour. Choose one of the

following:"

Answer options:

a) "Require it to be corrected; there are large, obvious errors",

b) "Require it to be corrected; there are minor errors",

c) "Accept it as it is; there are minor errors that need a small amount of editing",

d) "Accept it as it is; the contour is very precise".

Time evaluation

The time needed for the manual and automatic segmentation of every ROI for each patient, as well as the manual correction time of all ROIs, to be clinically acceptable, were recorded.

Results

Quantitative evaluation

For all 20 patients, the autocontouring module did not fail to generate contours for any Region of Interest (ROI). The results for every ROI, target or OAR are presented in groups according to the ROIs volume. Group A consists of ROIs smaller than 1 ml, group B consists of ROIs between 1 ml and 3 ml, Group C consists of ROIs between 3 and 20 ml, Group D consists of ROIs between 20 and 40 ml and group E consists of ROIs larger than 40 ml (Table 1).

Group A (<1 ml)	bilateral lenses, bilateral lacrimal glands, hypophysis, bilateral optic nerves, optic chiasma
Group B (1-3 ml)	bilateral inner ears, bilateral mandible (temporomandibular) joints, level Ia, hyoid
Group C (3-20 ml)	bilateral eyes, bilateral submandibular glands, bilateral level III, bilateral Level IV
Group D (20-40 ml)	spinal cord, brainstem, larynx, bilateral parotids, bilateral sternocleidomastoids, trachea, bilateral level Ib, bilateral level II, bilateral level V
Group E (>40 ml)	mandible, brain, body surface

For a more straightforward visual interpretation of the quantitative indexes (DICE, HD, VR), presented in the following graphs, a common colour pattern was used:

Color	Values Range	Interpretation
Red	DICE: 0.0-0.5, HD: 8-10mm, VR: <0.4 OR >1.6	Poor
Orange	DICE: 0.5-0.6, HD: 6-8mm, VR: 0.4-0.6 OR 1.4-1.6	Poor intermediate
Yellow	DICE: 0.6-0.7, HD: 4-6mm, VR: 0.6-0.8 OR 1.2-1.4	Good intermediate
Green	DICE: 0.7-1.0, HD: 0-4mm, VR: 0.8-1.2	Good

Table 1. Regions of Interest (ROIs) grouped according to their volume.

For Group A (<1 ml) the autocontouring module had a poor or poor intermediate performance regarding DICE for all ROIs (Figure 1). Bilateral lacrimal glands DICE values were the lowest, (Lacrimal Gland L: 0.27±0.07, Lacrimal Gland R: 0.29±0.10), while DICE values for bilateral optic nerves were the highest DICE values in Group A (Optic Nerve L: 0.56±0.11, Optic Nerve R: 0.56±0.12).

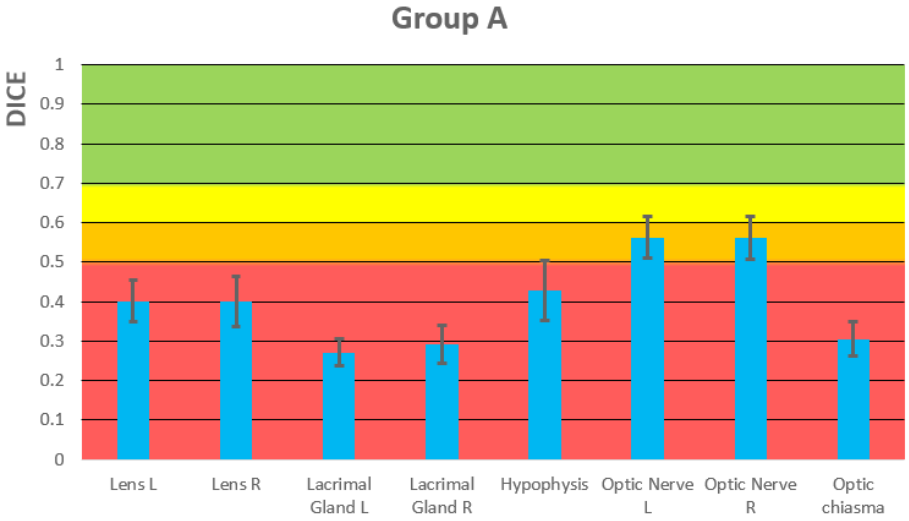


Figure 1: Average and 95% Confidence Interval of DICE for Group A.

Regarding HD values for group A, the autocontouring module had a good intermediate to good performance for all ROIs apart from optic chiasma (Figure 2). Bilateral optic nerves, lenses and hypophysis had the lowest HD values (Optic Nerve L: 3.47±1.34 mm, Optic Nerve R: 3.67±1.36 mm, Lens L: 3.88±1.68 mm, Lens R: 3.64±1.04 mm, Hypophysis: 3.73±1.37 mm), while optic chiasma had the highest HD values (6.43±1.95 mm) in group A.

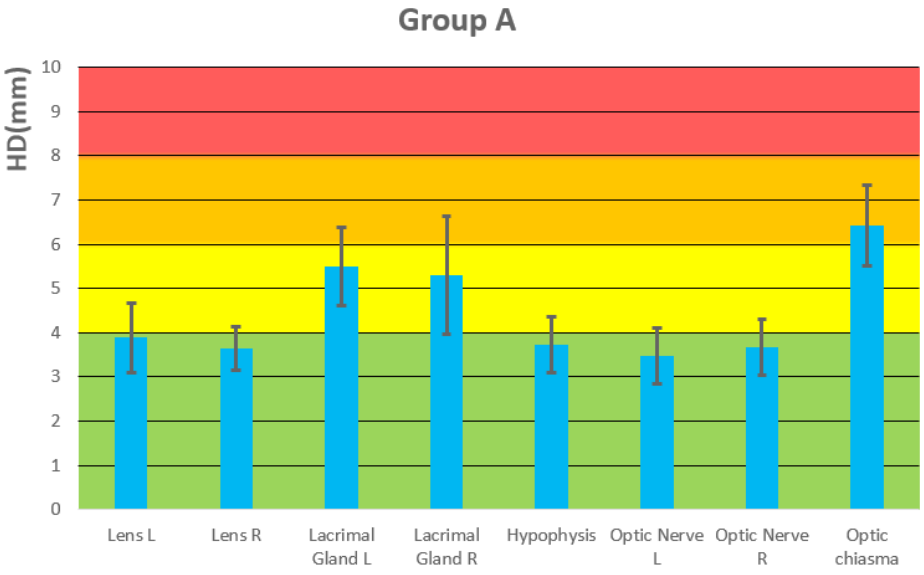


Figure 2: Average and 95% Confidence Interval of HD (95th percentile) for Group A.

Regarding VR values for group A, the autocontouring module had a good intermediate to good performance for all ROIs apart from bilateral lenses (poor intermediate to poor performance) (Figure 3). Bilateral optic nerves and optic chiasma had the best VR values (Optic Nerve L: 0.87±0.40, Optic Nerve R: 1.01±0.32, Optic chiasma: 0.93±0.40), while bilateral lenses had the worst VR values (Lens L: 1.54±1.10, Lens R: 1.91±1.48) in group A.

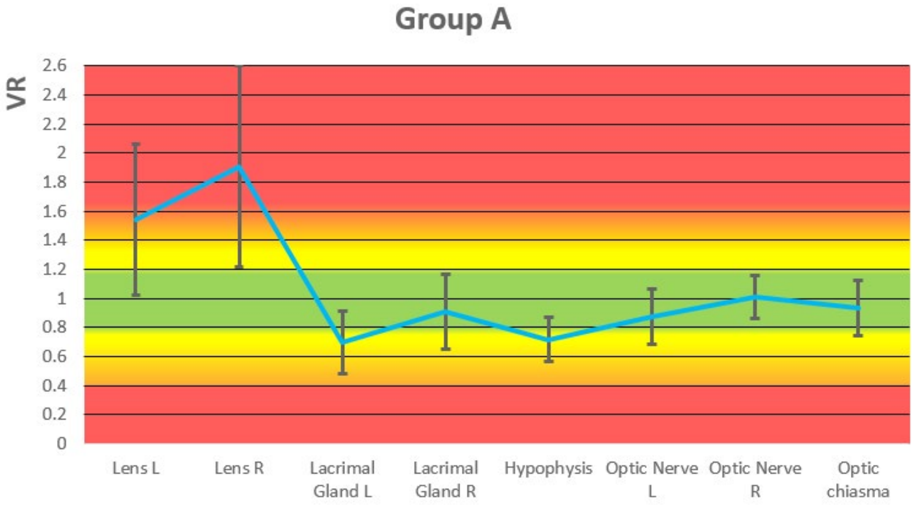


Figure 3: Average and 95% Confidence Interval of VR for Group A.

For Group B (1-3 ml) the autocontouring module had a poor intermediate to good intermediate performance regarding DICE (Figure 4) for all ROIs apart from Level IA (poor performance). Bilateral mandible (temporomandibular) joints DICE values were the highest (Mandible Joint L: 0.68±0.11, Mandible Joint R: 0.65±0.10), while the DICE values for Level IA were the lowest in Group B (Level IA: 0.40±0.14).

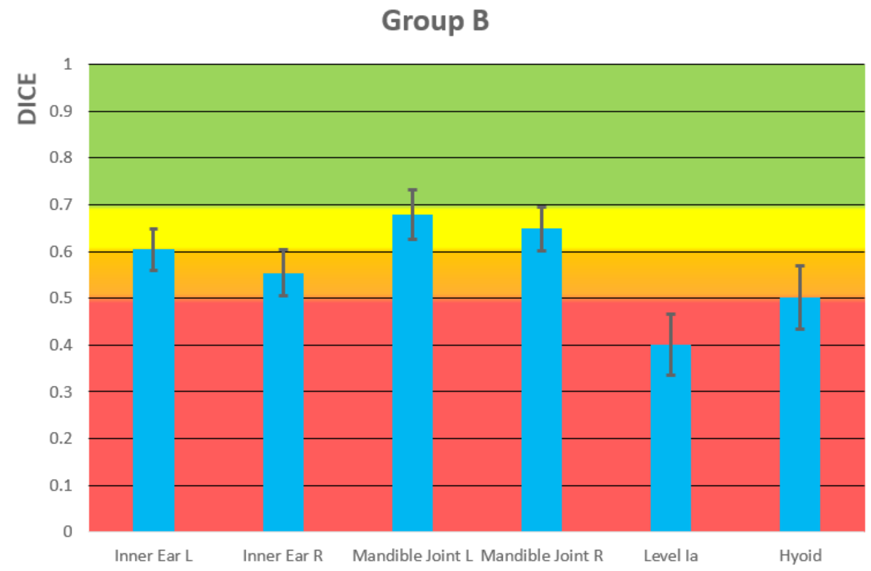


Figure 4: Average and 95% Confidence Interval of DICE for Group B.

Regarding HD values for group B, the autocontouring module had a good performance for all ROIs apart from Hyoid (poor intermediate performance) and Level IA (poor performance). (Figure 5). Bilateral inner ears and mandible (temporomandibular) joints had the lowest HD values (Inner Ear L: 3.73±0.97 mm, Inner Ear R: 4.15±1.53 mm, Mandible Joint L: 3.87±1.48 mm, Mandible Joint R: 4.01±1.58 mm), while Level IA had the highest HD value (8.68±4.38 mm) in group B.

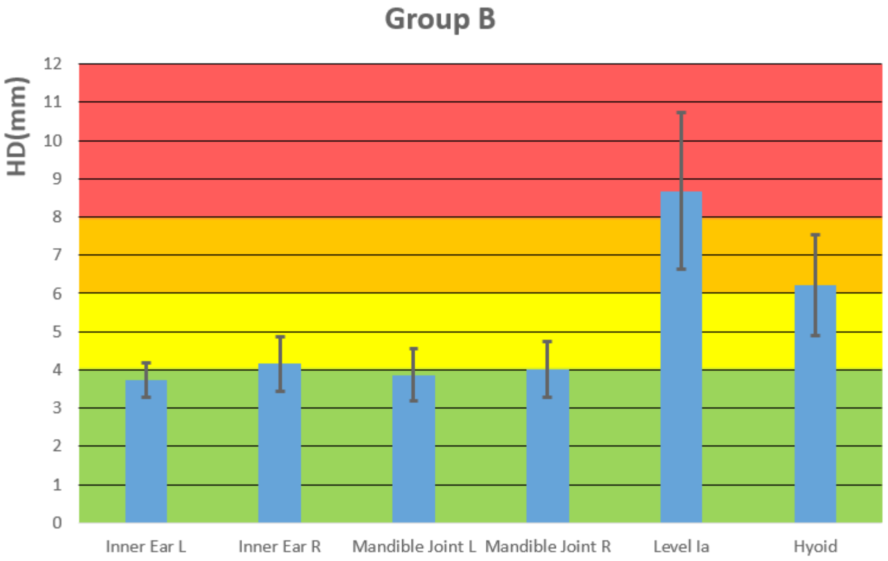


Figure 5: Average and 95% Confidence Interval of HD (95th percentile) for Group B.

Regarding VR values for group B, the autocontouring module had a good performance for bilateral mandible (temporomandibular) joints and hyoid, while for bilateral inner ears and Level IA (Figure 6), the performance was poor intermediate and poor respectively. Bilateral mandible joints and hyoid had the best VR values (Mandible Joint L: 1.11±0.46, Mandible Joint R: 1.14±0.43, Hyoid: 0.99±0.33), while Level IA had the worst VR values (Level IA: 0.40±0.14) in group B.

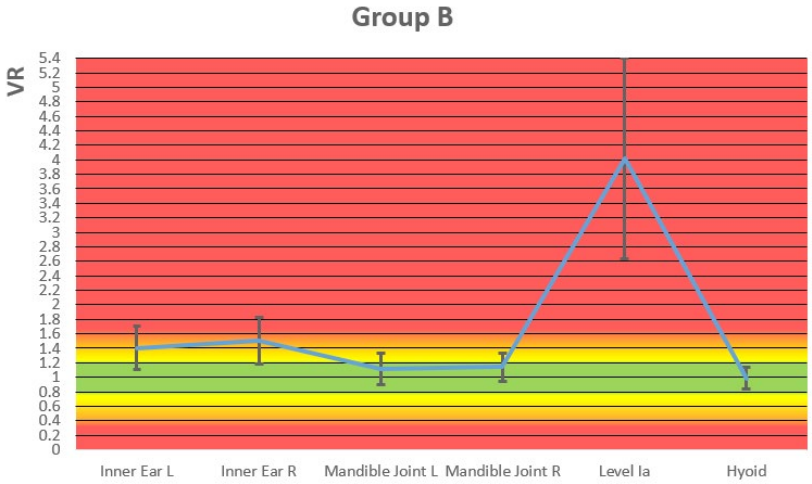


Figure 6: Average and 95% Confidence Interval of VR for Group B.

For Group C (3-20 ml) the autocontouring module had a good performance regarding DICE for bilateral eyes, a poor intermediate to good intermediate performance for bilateral submandibular glands and a poor intermediate performance for bilateral Level III and IV (Figure 7). All ROIs apart from Level IA (poor performance). Bilateral eyes DICE values were the highest (Eye L: 0.83±0.07, Eye R: 0.81±0.06), while the DICE values for bilateral Level IV were the lowest in Group C (Level IV L: 0.54±0.08, Level IV R: 0.50±0.09).

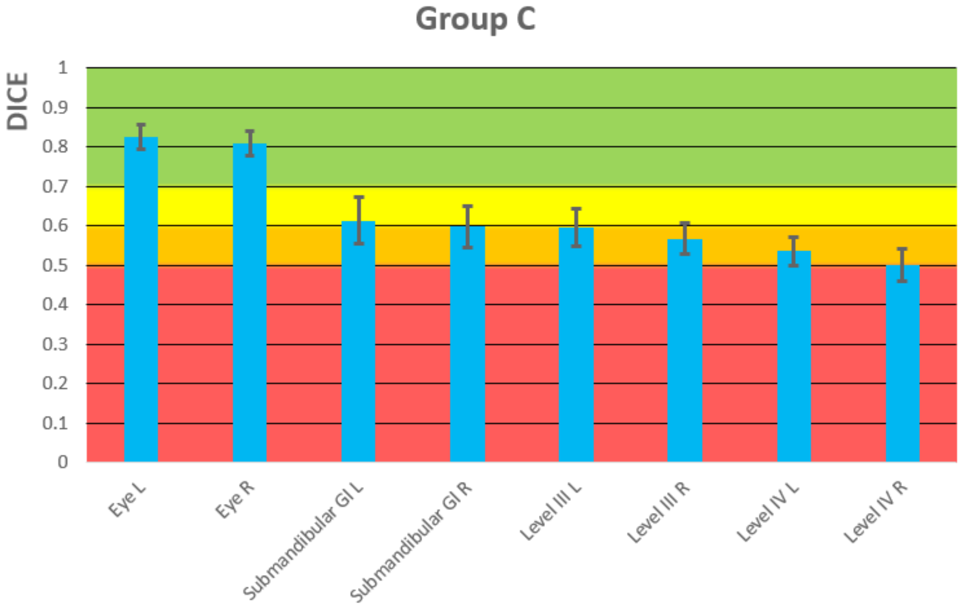


Figure 7: Average and 95% Confidence Interval of DICE for Group C.

Regarding HD values for group C, the autocontouring module had a good performance for bilateral eyes, a poor intermediate performance for bilateral submandibular glands and a poor performance for bilateral Level III and Level IV. (Figure 8). Bilateral eyes had the lowest HD values (Eye L: 3.30±0.99 mm, Eye R: 3.44±0.89 mm), while bilateral Levels III and IV had the highest HD values (Level III L: 9.45±4.43 mm, Level III R: 10.06±3.57 mm, Level IV L: 9.68±2.73 mm, Level IV R: 10.38±2.61 mm) in group C.

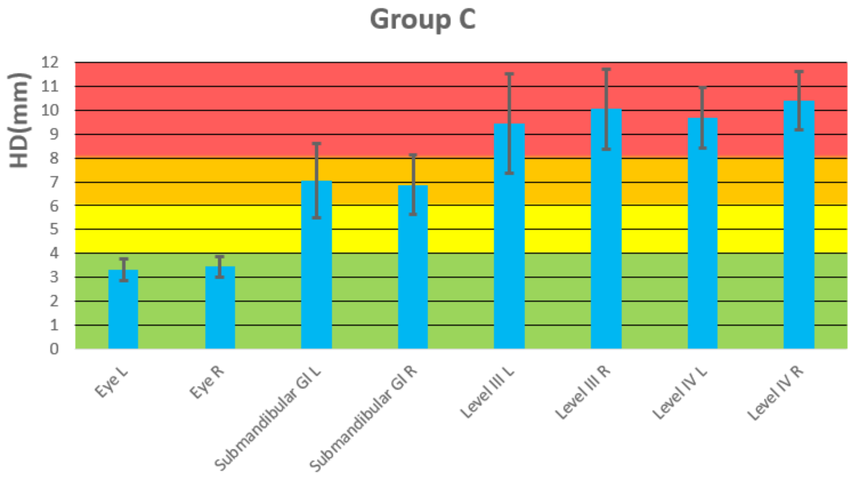


Figure 8: Average and 95% Confidence Interval of HD (95th percentile) for Group C.

Regarding VR values for group C, the autocontouring module had a good performance for all ROIs apart from bilateral Level IV (poor intermediate to poor performance) (Figure 9). Bilateral eyes and submandibular glands had the best VR values (Eye L: 0.84±0.27, Eye R: 0.85±0.26, Submandibular Gland L: 0.84±0.27, Submandibular Gland R: 0.82±0.34), while bilateral Level IV had the worst VR values (Level IV L: 1.35±0.44, Level IV R: 1.57±0.66) in group C.

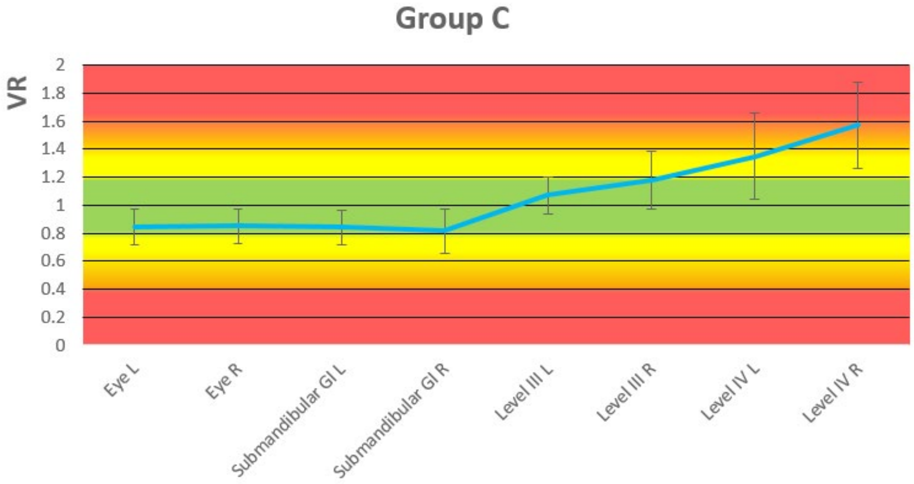


Figure 9: Average and 95% Confidence Interval of VR for Group C.

For Group D (20-40 ml) the autocontouring module had a good intermediate to good performance regarding DICE for all ROIs apart from bilateral Level V (poor intermediate performance) (Figure 10). Brainstem DICE value was the highest (0.76±0.07), while the DICE values for bilateral Level V were the lowest in Group D (Level V L: 0.51±0.15, Level V R: 0.50±0.12).

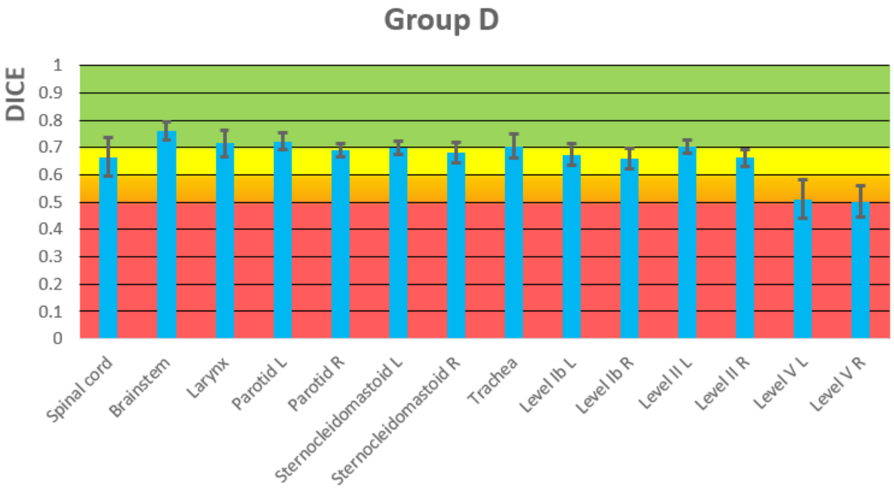


Figure 10: Average and 95% Confidence Interval of DICE for Group D.

Regarding HD values for group D, the autocontouring module had a poor performance for all ROIs apart from brainstem (good intermediate performance), right sternocleidomastoid and trachea (poor intermediate) (Figure 8). Brainstem had the lowest HD values (5.93±2.08 mm), while spinal cord had the highest HD values (15.08±18.60 mm) in group D.

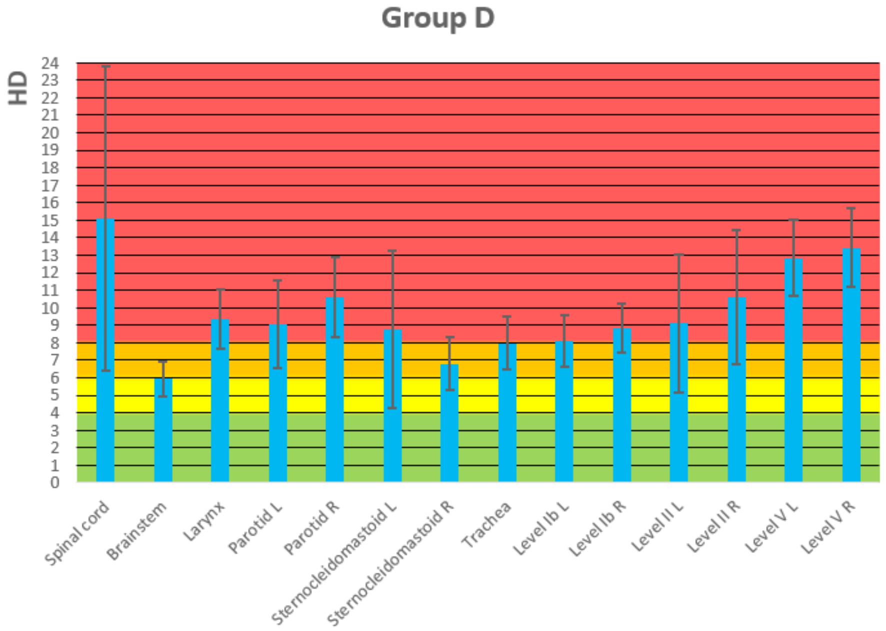


Figure 11: Average and 95% Confidence Interval of HD (95th percentile) for Group D.

Regarding VR values for group D, the autocontouring module had a good performance for all ROIs apart from right parotid and trachea (good intermediate performance) (Figure 12). Spinal cord had the best VR value (1.06±0.41), while trachea had the worst VR value (0.77±0.22) in group D.

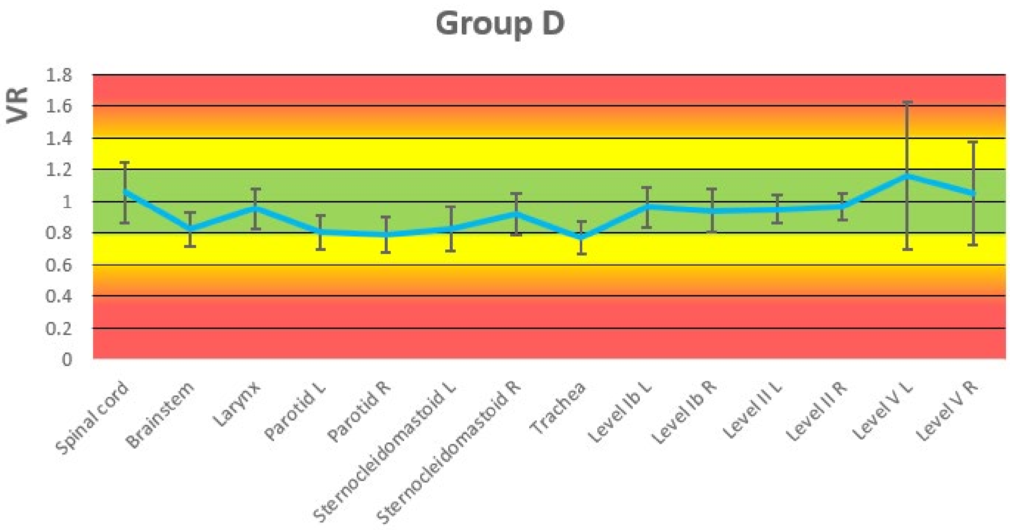


Figure 12: Average and 95% Confidence Interval of VR for Group D.

For Group E (>40ml) the autocontouring module had a good performance regarding DICE for all ROIs (Figure 13). Brain DICE value was the highest (0.97±0.01), while the DICE values for mandible was the lowest in Group E (0.79±0.09).

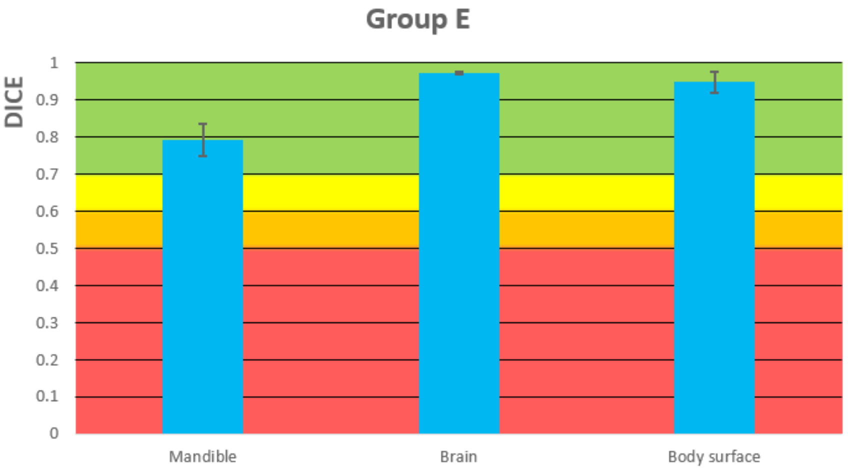


Figure 13: Average and 95% Confidence Interval of DICE for Group E.

Regarding HD values for group E, the autocontouring module had a good performance for brain, a good intermediate performance for mandible and a poor performance for body surface (Figure 14). Brain had the lowest HD values (3.97±2.55 mm), while body surface had the highest HD values (11.62±8.21 mm) in group E.

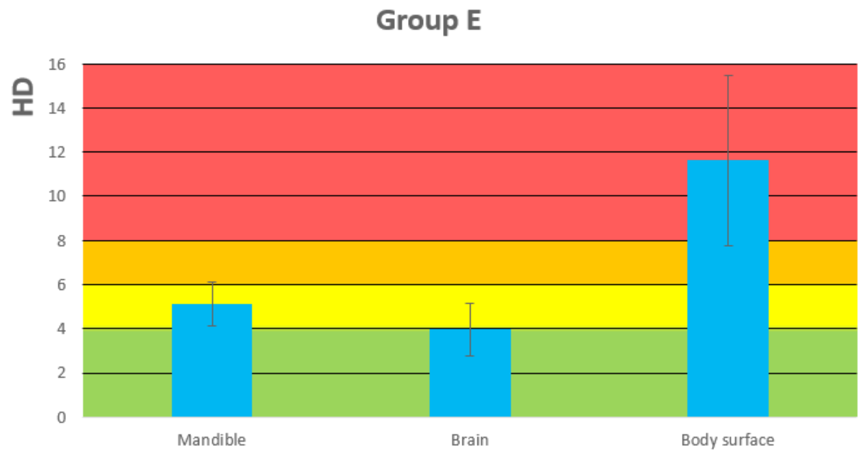


Figure 14: Average and 95% Confidence Interval of HD (95th percentile) for Group E.

Regarding VR values for group E, the autocontouring module had a good performance for all ROIs (Figure 15). Brain had the best VR value (1.00±0.03), while mandible had the worst VR value (1.12±0.17) in group E.

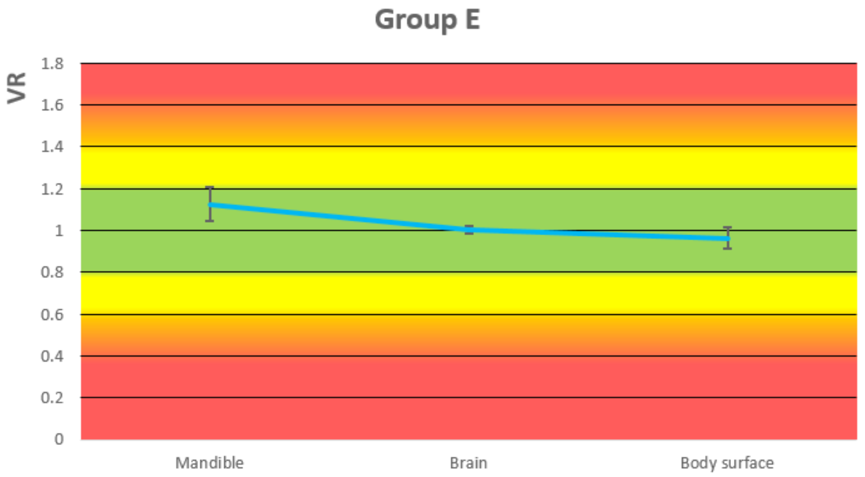


Figure 15: Average and 95% Confidence Interval of VR for Group D.

Subjective evaluation

For the question "whether contours were human or computer-drawn", 32% of the human-drawn contours were misclassified as computer-created. The misclassification rate for module's contours was 21% (Figure 16). The difference was greatest for bilateral lenses, bilateral lacrimal glands, chiasma and hypophysis.

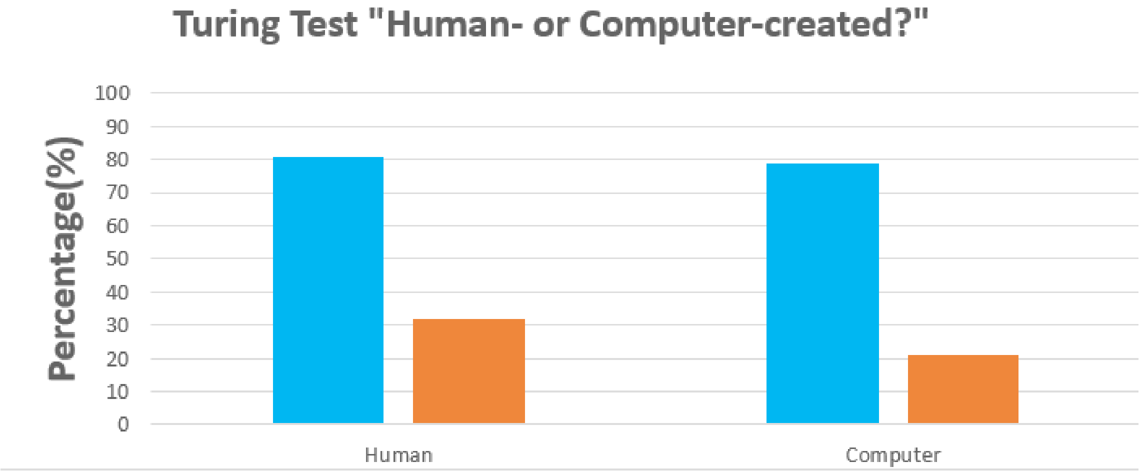


Figure 15: Percentage of right (blue) and wrong (orange) classification of the contours.

For the question "which of 2 contours was preferred", human-drawn contours were substantially more often preferred (81%) than either computer-created (19%). For almost all ROIs apart from brain, human-drawn contours were more often preferred (Figure 16).

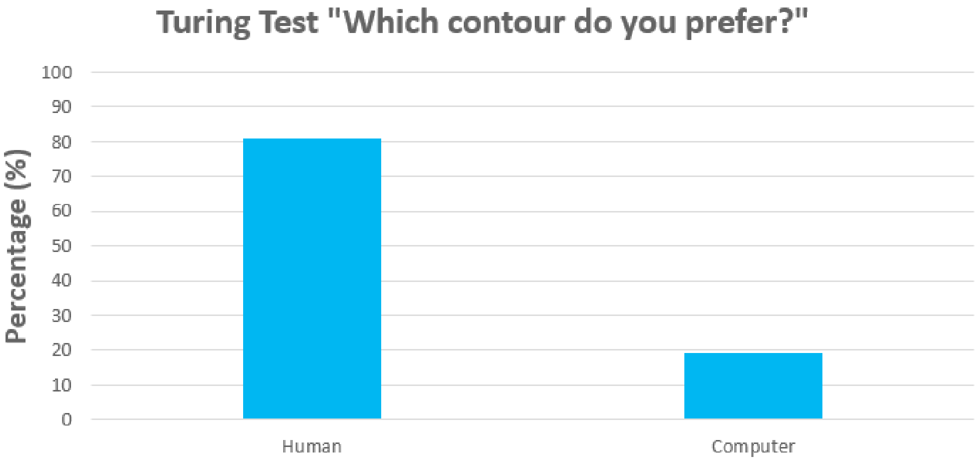
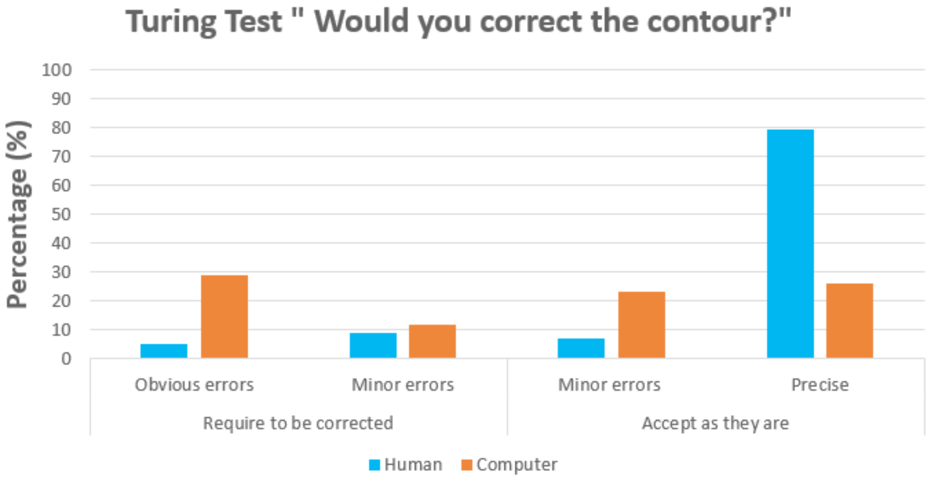


Figure 16: Percentage of contours preference.

The responses to the question "Would you correct the contour?" suggest low rates of large, obvious errors (5%) in human-drawn contours. In contrast, 29% of computer-created contours were considered to have obvious errors. Minor errors that required correction were less often found in human-drawn contours (9%) in comparison to computer-created contours (12%). 86% of the human-drawn contours and 59% of computer-created contours would be accepted would be accepted as they were with minor or without any corrections (Figure 17).



Time evaluation

The average time needed for the manual segmentation of each patient was 102,3 minutes, while the time needed for the autocontouring was 2,4 minutes. The average time needed for the manual correction of all ROIs for each patient in order to be clinically acceptable was 52,7 minutes.

Discussion

Target and OARs manual delineation is not only a challenging task that is time consuming, but also can lead to uncertainties due to inter- and intra-observer variations in modern radiotherapy that utilizes precise techniques. Additionally, this challenging task is constantly getting more complex, as an increasing number of OARs is shown to be associated with radiation-induced side effects, especially in the head and neck region. Against this background, auto-contouring of target volumes and OARs has the potential to reduce delineation precision, time and effort, and to improve inter-observer consistency [10-12].

Due to recent technological developments, deep learning techniques have emerged as a promising method of autodelineation. Training a Convolutional Neural Network (CNN) model directly from high-quality data has improved the autosegmentation results [18-22]. It seems that high quality data is the key for an efficient CNN. Manually edited ABAS-atlases, that can minimise the discrepancies among clinicians, have the potential to simplify the collection of such data, while respecting the personal contouring preferences of the clinicians.

In our study we evaluated a commercial ABAS module included into a dedicated radiotherapy contouring software (ProSoma, Medcom GmbH, Darmstadt, Germany), which is based on an atlas-library consisting of 100 patients. We assessed the performance of the module using 20 patients' DICOM files, delineating 39 ROIs, and comparing the manual contours to the module-created contours. Quantitative and subjective evaluation was performed. It has to be noted that the metrics we used, are common metrics used in comparing contours, but their interpretation is not always straightforward. As new technologies are tested, we strongly believe that standardised, reliable sets of both quantitative and subjective metrics will arise, allowing the more accurate and reproducible contours comparison methods. According to the authors' knowledge, our study is the study with the most different ROIs tested (39) regarding the head and neck region in the literature.

In the following section we will compare the results of our study with these of previous published studies, and discuss the reason for the discrepancies, if any.

Quantitative Evaluation

As a large number of studies have performed quantitative evaluation of the geometric accuracy of ABAS techniques in the head and neck region [28-40], using similar metrics (DICE, HD, VR), the results of our study will be compared to the findings of these for every ROI.

Lenses: Fortunati, et al. proposed an autosegmentation method primarily for hyperthermia treatment of head and neck tumours that combines anatomical information based on atlas registration with local intensity information in a graph-cut framework [28]. This novel method was compared with a multi-atlas ABAS method for multiple organs including lenses. The ABAS method they used provided median DICE and HD values of 0.41 and 4.96 mm, while the proposed novel method achieved median DICE and HD values of 0.67 and 3.73 mm. Our results are comparable to the ABAS methods used, with similar average DICE (Lens L:0.40±0.11 mm, Lens R: 0.40±0.13 mm) and HD values (Lens L:3.88±1.68 mm, Lens R: 3.64±1.04 mm).

Lacrimal Glands: As no studies were found in the literature that used ABAS methods for autosegmentation of the bilateral lacrimal glands, a recent study by Nikolov, et al. that used deep learning techniques is used for comparison [29].The authors reported DICE values of 0.62±0.13 and 0.69±0.12 for the left lens, as well as DICE values of 0.61±0.11 and 0.70±0.12 for the right lens. Our results (average DICE: 0.27±0.07 and 0.29±0.10 for the left and right lacrimal gland, respectively), are significantly worse. This can be explained by the fact that different methods were used. It is widely accepted that deep learning techniques provide better autosegmentation results [21,22].

Hypophysis: As no studies were found in the literature that used ABAS methods for autosegmentation of the hypophysis, a study by Tao, et al. that used an ABAS method followed by manual correction of the ROIs and compared the results to expert contours, to assess interobserver variability, is used for comparison [29]. Tao, et al. reported DICE indexes<0.41 for all small-volume ROIs, including hypophysis and lenses. Our results (average DICE: 0.43±0.16) are comparable to the aforementioned.

Optic Nerves: Walker, et al. used a commercial ABAS method for multiple head and neck ROIs, including optic nerves, and compared it to ABAS followed by manual editing and manual segmentation [30]. They showed DICE values of 0.71±0.26 for the ABAS method used. These DICE indexes are superior to our results (Optic Nerve L: 0.56±0.11, Optic Nerve R: 0.56±0.12). Possible explanations could be the different delineation extent of optic nerves and optic chiasma or the better performance of the ABAS method used for this specific ROI.

Optic Chiasma: Walker, et al. also report.ed DICE values for optic chiasma [35]. Their results (0.37±0.32) are similar to ours (0.31±0.01). It should be noted that other studies also report similarly low DICE values (0.39-0.57) regardless of the type of autocontouring method used, including deep learning methods [31-33].

Inner Ears: As no studies were found that used ABAS methods for autosegmentation of the bilateral inner ears, studies that used ABAS or deep learning methods for bilateral cochleae are used for comparison. Walker, et al. could show DICE values of 0.56±0.38 (ABAS) [30], while Thomson et al reported average DICE values of 0.30 [34]. It should be noted that deep learning techniques have shown superior results (DICE: 0.65±0.15-0.82±0.70).

Mandible (temporomandibular) Joints: No studies were found that evaluated any autocontouring method for the above ROIs, although some authors consider them significant OARs [35]. Our quantitative indexes have been presented previously.

Level IA-V: As no studies were found that used ABAS methods for autosegmentation of the head and neck lymphatic levels, separately, two studies that used ABAS for all the lymphatic levels as a single ROI are used for comparison. Yang et al. have shown DICE and HD indexes of 0.77±0.03 and 10±2.1mm for a single-atlas approach, as well as DICE and HD indexes 0.78±0.03 and 9.9±1.7mm [36] for a multi-atlas approach, respectively. Voet, et al. reported similar DICE values of 0.81 (0.69-0.95) and 0.82 (0.64-0.89) [37]. Although our indexes (average DICE for Levels IA-V:0.57) appear to be significantly inferior, the results are not comparable, as both studies compared the whole lymphatics as a single ROI.

Hyoid: No studies were found that evaluated any autocontouring method for the above ROI. Our quantitative indexes have been presented previously.

Eyes: Duc, et al. used two different ABAS methods for segmentation of bilateral eyes and compared the quantitative metrics to manual contours [31]. The authors reported DICE and HD values of <0.6 and 6-7 mm for both methods, respectively. Our results (DICE: Eye L: 0.83±0.07, Eye L: 0.81±0.06, HD: Eye L: 3.73±0.99, Eye R: 3.44±0.89) are significantly superior. Similar results to ours have been published by Zhu, et al. [38]. The authors reported DICE and HD values of 0.85±0.08 and <0.4 mm respectively.

Submandibular Glands: Walker, et al. reported DICE values for submandibular glands [30]. Their results (0.73±0.25) are slightly to ours (DICE: Submandibular L: 0.61±0.13: Submandibular R: 0.60±0.11). Other studies also report similar DICE values (0.65-0.80) [39,40].

Spinal Cord: Walker, et al. also reported DICE values for spinal cord [30]. The authors have published DICE values of 0.90±0.14. Other authors have reported slightly worse DICE and HD values of 0.6-0.75 and 0.40-0.45 mm. Our results (DICE: 0.67±0.15, HD:15±18 mm) are similar to the reported results in the literature apart from the HD index. This is primarily due to the non-standard length of the DICOM files used for the evaluation of the module.

Brainstem: Fortunati, et al. have also published DICE and HD values for the above ROI [28]. They have shown DICE and HD values of 0.78 and 9.07mm respectively. Similarly, Zhu et al. reported DICE and HD values of <0.8 and >5 mm respectively [38]. Our results (DICE: 0.76±0.07, HD: 5.93±2.08) are similar to the majority of the published studies [28,31,38]. Walker et al. have reported higher DICE values (0.97±0.03) utilising an ABAS method [30].

Larynx: Thomson, et al. evaluated an ABAS method, followed by manual editing for multiple head and neck ROIs, including larynx [34]. Their results (DICE: 0.76) are similar to ours (DICE: 0.72±0.11)

Parotids: Zhu, et al. have also reported DICE and HD values for bilateral parotids using an ABAS method [38]. Their values (DICE: 0.72±0.12, HD:>9 mm) are comparable to our results

(DICE: Parotid L: 0.72±0.06, Parotid R: 0.69±0.05, HD: Parotid L: 9.04±5.37 mm, Parotid R: 10.59±4.84 mm). Walker, et al. have reported higher DICE values (0.89±0.11) [30].

Sternocleidomastoids: No studies were found that evaluated any autocontouring method for the above ROI. Our quantitative indexes have been presented previously.

Trachea: No studies were found that evaluated any autocontouring method for the above ROI. Our quantitative indexes have been presented previously.

Mandible: Zhu, et al. have also reported DICE and HD values for mandible [38]. Their values (DICE: 0.85±0.04, HD:>9 mm) are similar to our results (DICE: 0.79±0.09, HD: 5.10±2.12 mm). Walker, et al. and others have reported slightly superior DICE values (0.89-0.91) [30,40].

Brain: Zhu, et al. have also published DICE and HD values for brain [38]. Their findings (DICE: 0.96±0.02, HD:<8 mm) are similar to ours (DICE: 0.97±0.01, HD: 3.97±2.55 mm).

Body surface: No studies were found that evaluated any autocontouring method for the above ROI. Our quantitative indexes have been presented previously.

It has to be noted that assessing clinical usability of contours based on geometric measures alone is very challenging. Moreover, comparisons between autocontouring methods based on metrics, such as DICE, HD or VR is of unknown value.

Subjective evaluation

The Turing test we performed, showed that although not all computer-drawn contours are precise, it is not always straightforward to identify contours as being human- or computer-created. The fact that 32% of the human-drawn contours were misclassified as computer-created and 21% of the computer-drawn contours were misclassified as human-created, indicates the degree intra- and interobserver variability. Although manual contours seem to be more often preferred, almost 60% of the computer-drawn contours would be accepted with minor or no editing.

A significant limitation of the Turing test we performed, is the presentation of single slices for ROIs evaluation. If the three-dimensional information was provided to the observer, the results could have been different.

Time Evaluation

The time evaluation suggests that ABAS, followed by manual editing, can significantly reduce the time and effort needed for the segmentation of the ROIs of the head and neck region. ABAS methods have been clinically implemented for adaptive radiation therapy for a number of years in many radiotherapy clinics and are being increasingly used in the daily routine. For institutions that wish to build neural networks, manually edited ABAS methods might be the procedure of choice.

Conclusion

It is self-explanatory that autocontouring will play an important role in the future of radiotherapy planning. The commercial ABAS module included into a dedicated radiotherapy contouring software (ProSoma, Medcom GmbH, Darmstadt, Germany) that we evaluated, has a similar performance to other commercial atlas-based solutions in terms of commonly used quantitative and subjective metrics. The time evaluation showed significant benefit for the daily routine. We demonstrated that using this module and fine-tuning the results is efficient for clinical utility and can considerably save time for clinicians in delineating ROIs for the head and neck region.

Conflict of Interest Statement

The authors would like to ensure that no financial support has been received in conjunction with the generation of the current submission and none of the authors has any personal or institutional financial interest in drugs or materials described in this paper.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, et al. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68: 394-424.
2. Parkin DM, Bray F, Ferlay J, Pisani P (2005) Global cancer statistics, 2002. CA Cancer J Clin 55: 74-108.
3. Zhang LW, Li J, Cong X, Hu XS, Li D, et al. (2018) Incidence and mortality trends in oral and oropharyngeal cancers in China, 2005-2013. Cancer Epidemiology. 57: 120-126.
4. Gupta B, Bray F, Kumar N, Johnson NW (2017) Associations between oral hygiene habits, diet, tobacco and alcohol and risk of oral cancer: A case–control study from India. Cancer Epidemiology 51: 7-14.
5. Zhao D, Xu QG, Chen XM, Fan MW (2009) Human Papillomavirus as an Independent Predictor in Oral Squamous Cell Cancer. International Journal of Oral Science 1: 119-125.
6. Mehrtash H, Duncan K, Parascandola M, David A, Gritz ER, et al. (2017) Defining a global research and policy agenda for betel quid and areca nut. Lancet Oncol 18: 767-775.
7. Eisbruch A, Ten Haken RK, Kim HM, Marsh LH, Ship JA (1999) Dose, volume, and function relationships in parotid salivary glands following conformal and intensity-modulated irradiation of head and neck cancer. Int J Radiat Oncol Biol Phys 45: 577–587.
8. Grégoire V, Evans M, Le QT, Bourhis J, Budach V, et al. (2018) Delineation of the primary tumour Clinical Target Volumes (CTV-P) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: AIRO, CACA, DAHANCA, EORTC, GEORCC, GORTEC, HKNPCSG, HNCIG, IAG-KHT, LPRHHT, NCIC CTG, NCRI, NRG Oncology, PHNS, SBRT, SOMERA, SRO, SSHNO, TROG consensus guidelines. Radiother Oncol 126: 3-24.

9. Hall WH, Guiou M, Lee NY, Dublin A, Narayan S, et al. (2018) Development and Validation of a Standardized Method for Contouring the Brachial Plexus: Preliminary Dosimetric Analysis Among Patients Treated With IMRT for Head-and-Neck Cancer. International Journal of Radiation Oncology Biology Physics 72: 1362–1367.

10. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, et al. (2014) Vision 20/20: Perspectives on automated image segmentation for radiotherapy. Medical Physics 41: 1-14.

11. Lim JY, Leech M (2016) Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. Acta Oncologica 55: 799-806.

12. Valentini V, Boldrini L, Damiani A, Muren LP (2014) Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. Radiother Oncol 112: 317–320.

13. Rohlfing T, Brandt R, Menzel R, Russakoff DB (2007) Quo Vadis, Atlas-based segmentation?. In: Suri JS, Wilson DL, Laxminarayan S, editors. Handbook of Biomedical Image Analysis: Volume 2: Segmentation Models Part B. 2005 ed. New York: PC Springer: p. 435-486.

14. Larrue A, Gujral D, Nutting C, Gooding M (2015) The impact of the number of atlases on the performance of automatic multi-atlas contouring. Phys Med 31: e23-e54.

15. Yeo UJ, Supple JR, Taylor ML, Smith R, Kron T, et al. (2013) Perfomance of 12 DIR algorithms in low-contrast regions for mass and density conserving deformation. Med Phys 40: 1–12.

16. Zhong H, Kim J, Chetty IJ (2010) Analysis of deformable image registration accuracy using computational modeling. Med Phys 37: 970–979.

17. Kim H, Jung J, Kim J, Cho B, Kwak J, et al. (2020) Abdominal multi-organ auto-segmentation using 3D-patch-based deep convolutional neural network. Sci Rep 10: 1-9.

18. Ibragimov B, Xing L (2017) Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys 44: 547–557.

19. Liang S, Tang F, Huang X, Yang K, Zhong T, et al. (2018) Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. Eur Radiol 28: 1-7.

20. Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, et al. (2018) Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiother Oncol 126: 312–317.

21. Men K, Dai J, Li Y (2017) Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks. Med Phys 44: 6377–6389.

22. Ibragimov B, Toesca D, Chang D, Koong A, Xing L (2017) Combining deep learning with anatomical analysis for segmentation of the portal vein for liver SBRT planning. Phys Med Biol 62: 8943–8958.

23. Gregoire V, Ang K, Budach W, Grau C, Hamoir M, et al. (2014) Delineation of the neck node levels for head and neck tumors: A 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. Radiotherapy and Oncology 110: 172-181.

24. Brouwer CL, Steenbakkers RJ, Bourhis J, Budach W, Grau C, et al. (2015) CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. Radiother Oncol 117: 83-90.

25. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26: 297–302.

26. Heimann T, Van Ginneken B, Styner MA, Arzhaeva Y, Aurich V, et al. (2003) Comparison and evaluation of methods for liver segmentation from CT datasets. IEEE Trans Med Imaging 28: 280.

27. Gooding MJ, Smith AJ, Tariq M, Aljabar P, Peressutti D, et al. (2018) Comparative evaluation of auto-contouring in clinical practice: a practical method using the Turing Test. Med Phys 45: 1–11.

28. Fortunati V, Van Der Lijn F, Niessen WJ, Veenland JF, Paulides MM, et al. (2013) Tissue segmentation of head and neck CT images for treatment planning: a multiatlas approach combined with intensity modeling. Med Phys. 40: 71905–1–14.

29. Tao C, Yi J, Chen N, Ren W, Cheng J, et al. (2015) Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: a multi-institution clinical study. Radiother Oncol 115: 407–411.

30. Walker G, Awan M, Tao R, Koay E, Boehling N, et al. (2015) Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. Radiother Oncol 112: 321–325.

31. Hoang Duc AK, Eminowicz G, Mendes R, Wong S-L, Mcclelland J, et al. (2015) Validation of clinical acceptability of an atlas-based segmentation algorithm for the delineation of organs at risk in head and neck cancer. Med Phys 42: 5027–5034.

32. Ibragimov B, Xing L (2017) Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys 44: 547–557.

33. Menze BH, Jakab A, Bauer S, et al. (2015) The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans Med Imaging 34: 1993–2024.

34. Thomson D, Boylan C, Liptrot T, Aitkenhead A, Lee L, et al. (2014) Evaluation of an automatic segmentation algorithm for definition of head and neck organs at risk. Radiat Oncol 9: 1-12.

35. Goldstein M, Maxymiw WG, Cummings BJ, Wood RE (1999) The effects of antitumor irradiation on mandibular opening and mobility: a prospective study of 58 patients. Oral Surg Oral Med Oral Pathol Oral Radiol Endod 88: 365-373.

36. Yang J, Beadle BM, Garden AS, Gunn B, Rosenthal D, et al. (2014) Auto- segmentation of low-risk clinical target volume for head and neck radiation therapy. Pract Radiat Oncol 4: e31–37.

37. Voet PWJ, Dirkx MLP, Teguh DN, Hoogeman MS, Levendag PC, et al. (2011) Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage ? A dosimetric analysis. Radiother Oncol 98: 373–377.

38. Zhu M, Bzdusek K, Brink C, Eriksen JG, Hansen O, et al. (2013) Multi-institutional quantitative evaluation and clinical validation of Smart Probabilistic Image Contouring Engine (SPICE) autosegmentation of target structures and normal tissues on computer tomography images in the head and neck, thorax, liver, and male. Int J Radiat Oncol Biol Phys 87: 809–816.

39. Fritscher K, Raudaschl P, Zaffino P, Spadea MF, Sharp GC, et al. (2016) Deep neural networks for fast segmentation of 3D medical images," in Med Image Comput Comput Assist Interv. Springer International Publishing: 158–165.

40. Zhu W, Huang Y, Xie X, Zeng L, Chen X, et al. (2019) AnatomyNet: Deep 3D squeeze-and-excitation U-Nets for fast and fully automated whole-volume anatomical segmentation. Medical Phys 46: 576-589.